ACADEMIA DE STUDII ECONOMICE

Facultatea: Cibernetică, Statistică și Informatică Economică

Specializarea: Cibernetică Economică



**Proiect Analiza Datelor**

Studenți: Ivasișin Andrei, Murgeanu Andrei

Grupa: 1062

Profesor îndrumător: Matei Monica Mihaela

**Analiza Cluster**

Pentru Analiza Cluster am folosit datele standardizate pe care le-am utilizat si la primul proiect.

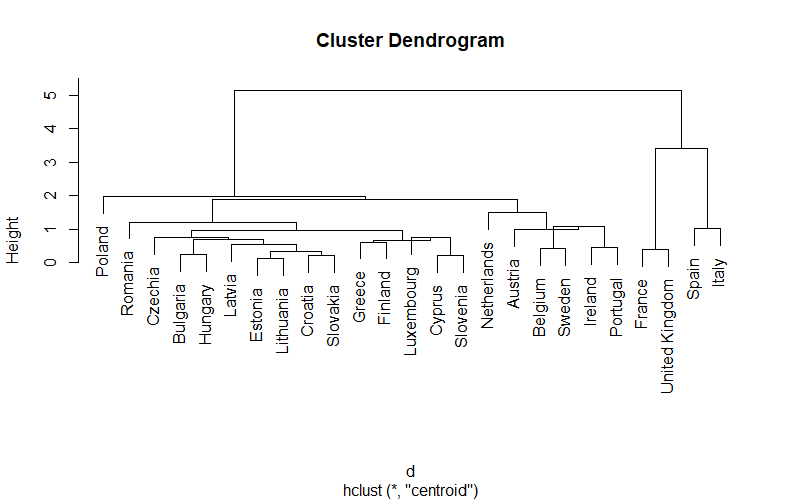
**Analiza cluster folosind metoda CENTROID**

Distantele de agregare dintre tarile alese sunt reprezentate in tabelul urmator:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 0.1213700 | 0.2168050 | 0.2186337 | 0.2374967 | 0.3395045 | 0.3814964 |
| 0.4113762 | 0.4429113 | 0.5497690 | 0.6153749 | 0.6892364 | 0.7376444 |
| 0.7592100 | 0.6699737 | 0.9474148 | 1.0187514 | 1.0848237 | 0.9904400 |
| 1.2124455 | 1.4907374 | 1.8909066 | 1.9740983 | 3.4037879 | 5.1435178 |

Prima grupare dintre Estonia si Lituania inregistreaza o distanta de 0.12137.

Cea de a doua grupare, dintre Slovenia si Cipru inregistreaza o distanta de 0.2168050

Putem observa diferente mari inregistrate intre 3.40 si 5.14, dar si diferente medii precum 1.49-1.89 sau 0.12-0.21.

Din dendrograma observam clasele care s-au realizat intre tarile care la momentul respectiv erau foarte asemanatoare. De exemplu: Bulgaria cu Ungaria, Estonia cu Lituania, Croatia cu Slovacia, Belgia cu Suedia, Spania cu Italia, Franta cu Regatul Unit, Irlanda cu Portugalia, etc.

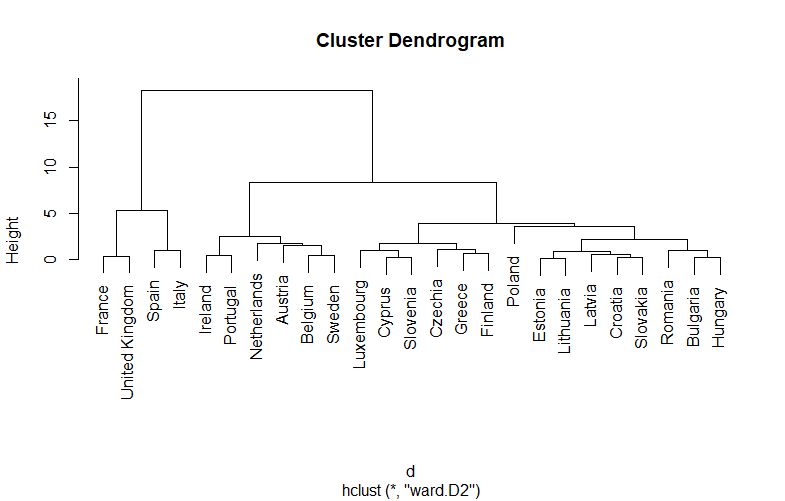
Distantele dintre clase s-au calculat pe baza centroizilor, intrucat mai intai s-au calculat centroizii celor doua clase, iar pe urma distanta dintre acestia.

Daca facem taietura sus , intre valorile 3-5, obtinem doua clase. Prima contine tarile Polonia,Romania,Cehia,Bulgaria,Ungaria,Letonia,Estonia,Lituania,Croatia,Slovacia,Grecia,Finlanda,Luxemburg,Cipru, Slovenia,Olanda,Austria,Belgia,Suedia,Irlanda si Portugalia, iar a doua clasa contine Franta, UK, Spania si Italia.

**Analiza cluster folosind metoda Ward**

Distantele de agregare dintre tarile alese sunt reprezentate in tabelul urmator:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 0.1213700 | 0.2168050 | 0.2186337 | 0.2374967 | 0.3814964 | 0.4113762 |
| 0.4429113 | 0.5601334 | 0.6153749 | 0.8728195 | 0.9389166 | 0.9746842 |
| 1.0187514 | 1.0643542 | 1.5113953 | 1.7172418 | 1.7244837 | 2.1265959 |
| 2.4836653 | 3.5355173 | 3.9538676 | 5.3005065 | 8.3648559 | 18.2449774 |



Din dendrograma observam clasele care s-au realizat intre tarile care la momentul respectiv erau foarte asemanatoare. De exemplu: Franta cu UK, Spania cu Italia, Irlanda cu portugalia, Belgia cu Suedia, Cipru cu Slovacia, Grecia cu Finlanda, Estonia cu lituania, Letonia cu Coratia, Bulgaria cu Ungaria.

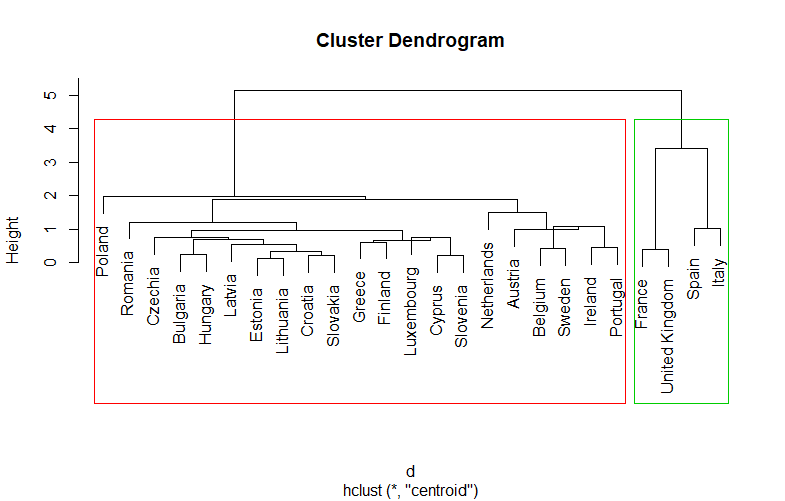
Daca facem taietura sus , intre valorile 10-15, obtinem doua clase. Prima contine tarile Franta, UK, Spania si Italia, iar a doua clasa contine Irlanda, Portugalia, Olanda, Austria,Belgia, pana la Ungaria.

Daca facem taietura mai jos, intre valoriale 5 si 10, obtinem trei clase.Prima se pastreaza si avem doua alte clase:

-prima: Irlanda, Portugalia, Olanda, Austria, Belgia, Suedia

- iar a doua clasa:Luxemburg, Cipru, Slovenia, Cehia, Grecia, Finlanda, Polonia, Estonia, Lituania, Letonia, Croatia, Slovacia, Romania, Bulgaria si Ungaria.

|  |  |
| --- | --- |
| Tara | Clasa |
| Belgia | 1 |
| Bulgaria | 1 |
| Cehia | 1 |
| Estonia | 1 |
| Irlanda | 1 |
| Grecia | 1 |
| Spania | 2 |
| Franta | 2 |
| Croatia | 1 |
| Italia | 2 |
| Cipru | 1 |
| Letonia | 1 |
| Lituania | 1 |
| Luxemburg | 1 |
| Ungaria | 1 |

**Solutia pentru analiza cluster cu metoda CENTROID**

Frecventa tarilor repartizate in cele doua clase este urmatoarea:

|  |  |
| --- | --- |
| 1 | 2 |
| 21 | 4 |

-21 de tari au fost repartizate in prima clasa

-4 tari au fost repartizate in a doua clasa

-nu este ecilibrat

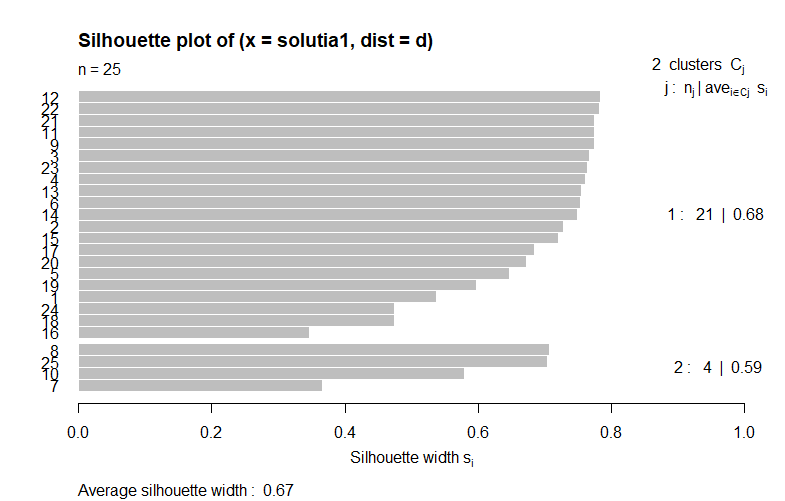
Din tabelul alaturat, putem observa urmatoarele:

Belgia, Bulgaria, Cehia, Estonia, Irlanda, Grecia, Croatia, Cipru, Letonia, Lituania, Luxemburg, Ungaria, etc sunt in prima clasa.

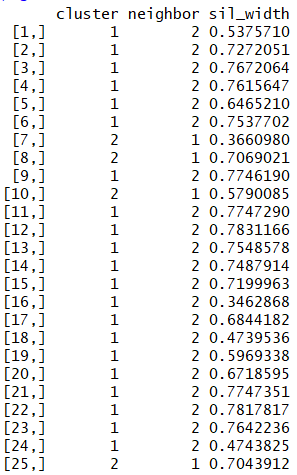
Franta, Spania, Italia, UK sunt in clasa a II-a.

**Silhouette plot**

Pentru a putea compara clasele intre ele , sa vedem care ar fi mai buna, utilizam silhouette plot.



Media silhouette este de 0.67.

Nu avem valori negative.

Inseamna ca toate variabile sunt repartizate corespunzator.

|  |  |  |
| --- | --- | --- |
| Group | Componenta 1 | Componenta 2 |
| 1 | -1.124391 | -0.06210326 |
| 2 | 5.903051 | 0.32604213 |

Pe linii avem cele doua clase, iar pe coloane cele doua caracteristici din analiza componentelor principale.

Media componentei 1 in prima clasa este -1.12.

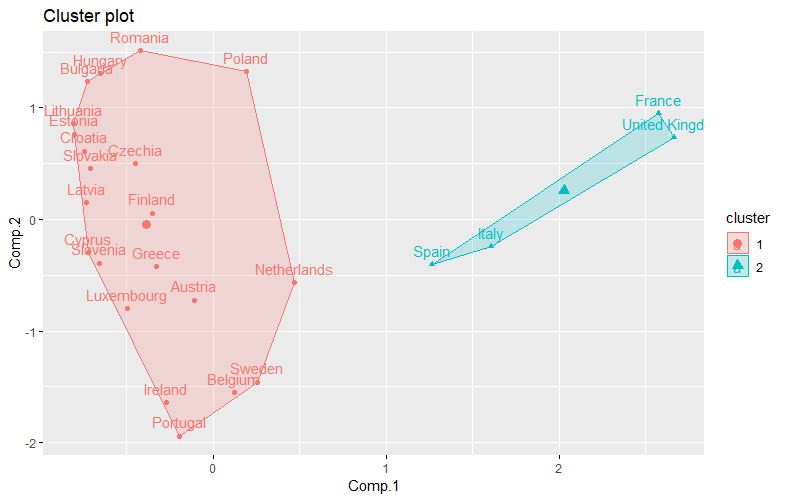
Media componentei 1 in a doua clasa este 5.9.

Media componentei 2 in prima clasa este -0.06.

Media componentei 2 in clasa a doua este 0.32.

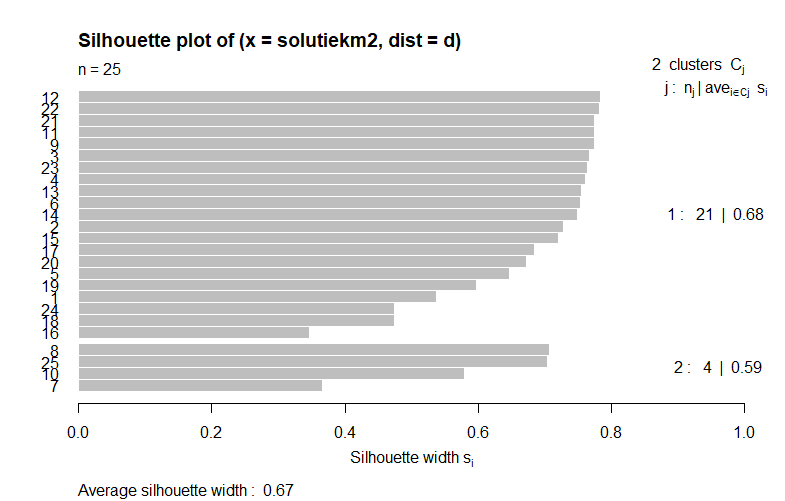
Observam ca prima componenta are valori mai mari (1.12 fata de 0.06 si 5.9 fata de 0.3), ceea ce inseamna ca a contribuit mai mult, intrucat aceasta este componenta care a preluat maximul de informatii cand am realizat analiza componentelor principale.

**Cluster Plot**



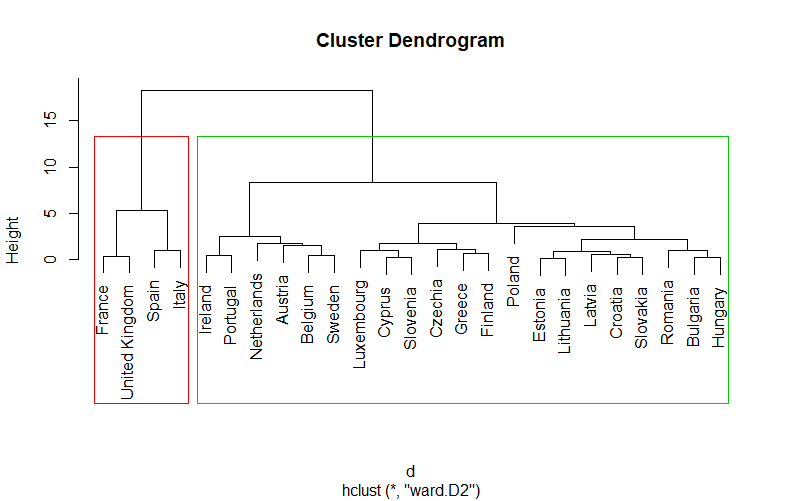
Observam cele doua clase formate si tarile care apartin lor. Clasa reprezentata cu rosu este puternic corelata cu a doua componenta, in ea regasindu-se tari care au un pret al combustibilui foarte ridicat( spre exemplu Romania,Polonia), in timp ce clasa reprezentata cu albastru este mult mai bine corelata cu prima componenta, deoarece in aceasta se regasesc tarile care au un numar mare de populatie, un salariu minim ridicat, un PIB mare si o productie de energie ridicata. Totodata tari precum Franta si UK sunt reprezentate si de un pret ridicat al combustibilului.

**Comparearea cu Kmeans**



Am obtinut silhouette cu aceeasi medie, in valoare de 0.67.

**Solutia pentru analiza cluster cu metoda WARD**

****

Frecventa tarilor repartizate in cele doua clase este urmatoarea:

|  |  |
| --- | --- |
| 1 | 2 |
| 21 | 4 |

-21 de tari au fost repartizate in prima clasa

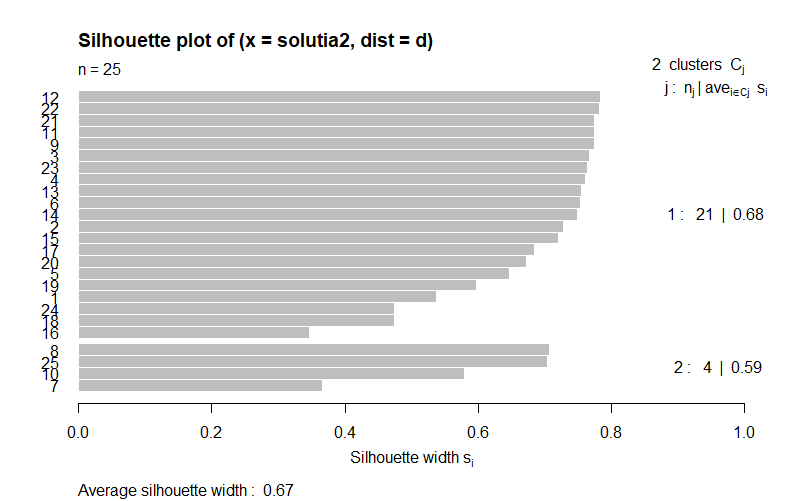
-4 tari au fost repartizate in a doua clasa

-nu este ecilibrat

Din tabelul alaturat, putem observa urmatoarele:

Belgia, Bulgaria, Cehia, Estonia, Irlanda, Grecia, Croatia, Cipru, Letonia, Lituania, Luxemburg, Ungaria, etc sunt in prima clasa.

Franta, Spania, Italia, UK sunt in clasa a II-a.



Media silhouette este de 0.67.

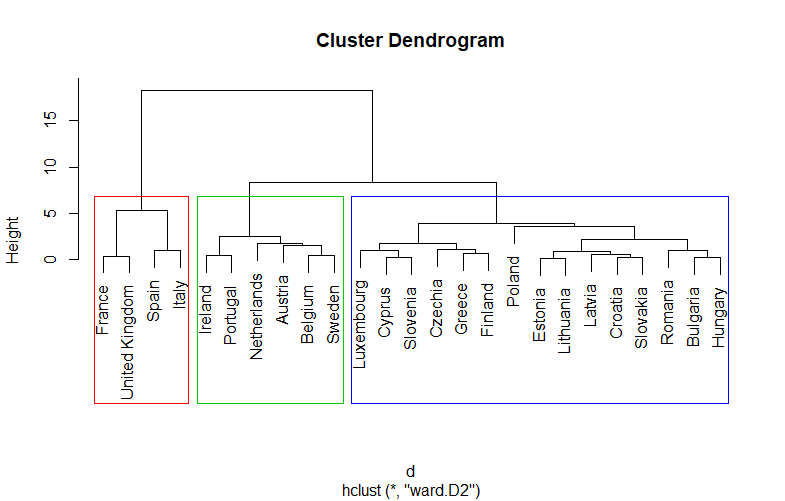
**OBSERVATIE:**

Rezultatul este acelasi ca in cazul metodei centroid, daca folosim prima taietura.

Vom face analiza cluster cu metoda Ward, utilizand taietura intre valorile 5-10.

|  |  |
| --- | --- |
| Tara | Clasa |
| Belgia | 1 |
| Bulgaria | 2 |
| Cehia | 2 |
| Estonia | 2 |
| Irlanda | 1 |
| Grecia | 2 |
| Spania | 3 |
| Franta | 3 |
| Croatia | 2 |
| Italia | 3 |
| Ciprua | 2 |
| Letonia | 2 |
| Lituania | 2 |
| Luxemb. | 2 |
| Olanda | 1 |

**Solutia pentru analiza cluster cu metoda WARD cu 3 clase**



Frecventa tarilor repartizate in cele trei clase este urmatoarea:

|  |  |  |
| --- | --- | --- |
| 1 | 2 | 3 |
| 6 | 15 | 4 |

-6 tari au fost repartizate in prima clasa

-15 tari au fost repartizate in a doua clasa

-4 tari au fost repartizate in a treia clasa

-cele 3 clase nu sunt echilibrate, intrucat cea de a doua este formata din 15

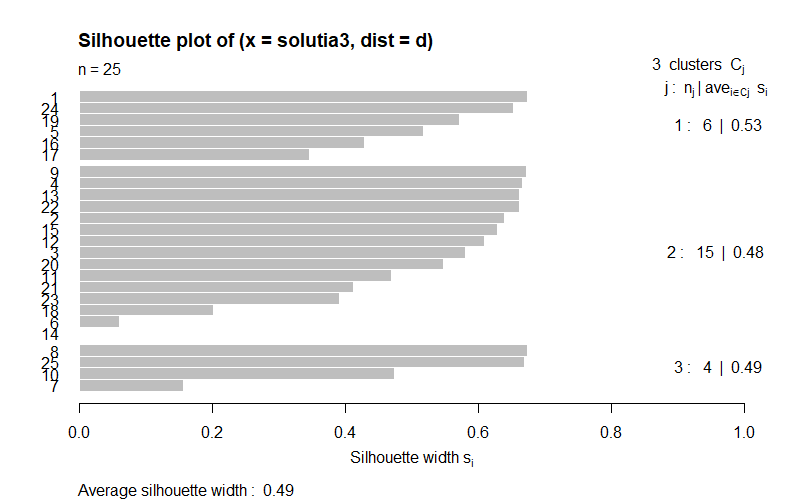
Din tabelul alaturat, putem observa urmatoarele:

Belgia, Irlanda, Olanda, etc fac parte din prima clasa.

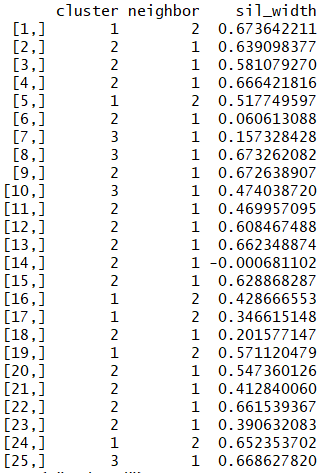
Bulgaria, Cehia, Estonia, Grecia, Croatia, Cipru, Letonia, Lituania, Luxemburg sunt in a II-a clasa.

A III-a clasa ramane neschimbata, continand tarile Spania, Franta, Italia, UK.

**Silhouette plot**



Media silhouette este de 0.49.



Avem o singura valoare negativa care este aproximativ 0. (-0.00068)

|  |  |  |
| --- | --- | --- |
| Grupa | Componenta 1 | Componenta 2 |
| 1 | 0.1346844 | -1.6682350 |
| 2 | -1.6280207 | 0.5803494 |
| 3 | 5.9030511 | 0.3260421 |

Pe linii avem cele trei clase, iar pe coloane cele doua caracteristici din analiza componentelor principale.

Media componentei 1 in prima clasa este 0.13.

Media componentei 1 in a doua clasa este -1.62.

Media componentei 1 in a treia clasa este 5.9.

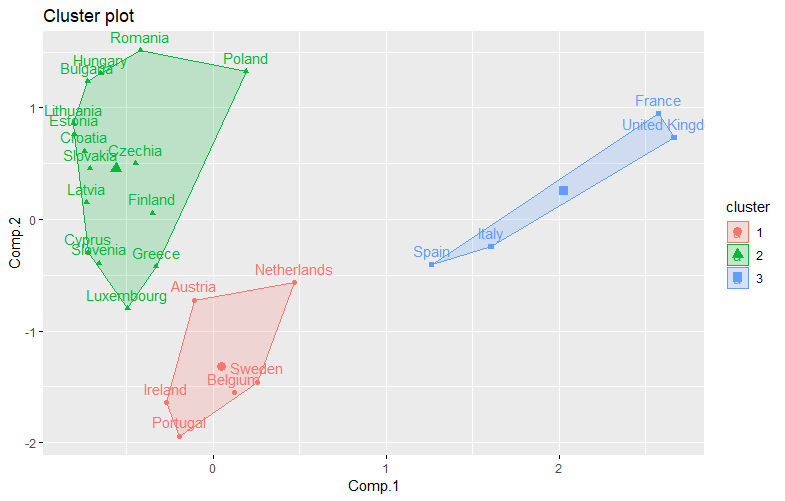
Media componentei 2 in prima clasa este -1.66.

Media componentei 2 in clasa a doua este 0.58.

Media componentei 2 in clasa a treia este 0.32.

Observam ca prima componenta are valori mai mari in grupele 2 si 3,ceea ce inseamna ca a contribuit mai mult, intrucat aceasta este componenta care a preluat maximul de informatii cand am realizat analiza componentelor principale. In schimb, componenta 2 are valori mai mari in clasa I, deoarece aceasta este componenta care caracterizeaza tarile cu un pret ridicat al combustibilului.

**Cluster Plot**

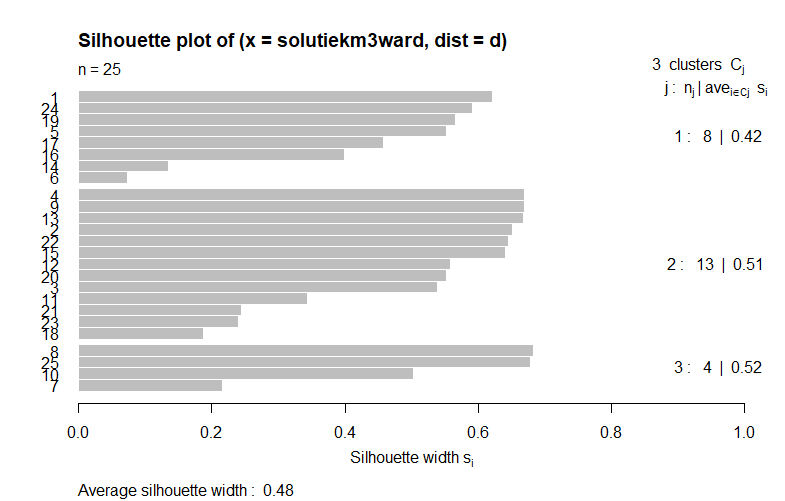


Din cluster plot observam usor cele 3 clase formate si tarile din care sunt formate.

Clasa reprezentata cu verde este puternic corelata cu a doua componenta, in ea regasindu-se tari care au un pret al combustibilui foarte ridicat, in timp ce clasa reprezentata cu albastru este mult mai bine corelata cu prima componenta, deoarece in aceasta se regasesc tarile care au un numar mare de populatie, un salariu minim ridicat, un PIB mare si o productie de energie ridicata. Totodata tari precum Franta si UK sunt reprezentate si de un pret ridicat al combustibilului.

In clasa marcata cu rosu se identifica tari cu un pret scazut al combustibilului si cu populatie,pib, salariu minim, productie de energie medii.

**Comparearea cu Kmeans**



Am obtinut silhouette cu media mai mica fata de cea precedenta (0.49), in valoare de 0.48; si fara valori negative.

Clasele sunt aproximativ echilibrate (0.42,0.51,0.52)

**Concluzie:**

In urma analizei efectuate, am obtinut trei outputuri prin intermediul Silhouette plot-ului.

Mediile Silhouette, la care am precizat si daca exista valori negative, sunt urmatoarele:

Metoda Centroid cu 2 clustere: 0.67, fara valori negative.

Metoda Ward cu 2 clustere: 0.67, fara valori negative.

Metoda Ward cu 3 clustere: 0.49, cu o singura valoare negativa aproximata la 0.

Centroid- Kmeans: 0.67, fara valori negative.

Ward- Kmeans: 0.48, fara valori negative.

Metoda cea mai buna este cea care are o medie Silhouette cat mai mare si in care nu se regasesc valori negative. Prin urmare, aceasta este reprezentata de metoda Centroid cu 2 clase.

Variabilitatile celei mai bune metode sunt urmatoarele:

Variabilitatea totala a acesteia este de 242.05

Variabilitatea in prima clasa este de 14.63, iar in a doua clasa este 60.97

Variabilitatea totala intraclasa este 75.61.

Variabilitatea interclasa este de 166.43.

|  |  |
| --- | --- |
| Componenta 1 | Componenta 2 |
| 5.903051 | 0.32604213 |
| - 1.124391 | - 0.06210326 |

ANEXA

path<-"C:\\Users\\Andrei\\Desktop"

date<-read.table(file.path(path,"AnalizaDatelor\_Proiect\_1.txt"),sep="\t",dec=".",header=TRUE,row.names = 1)

attach(date)

dates<-scale(date[-4,],center=TRUE,scale=TRUE)

acp<-princomp(dates,cor=TRUE,scores = TRUE)

m<-acp$scores[,1:2]

d<-dist(m)

#CENTROID

ierarhia1<-hclust(d,method="centroid")

ierarhia1$height #distanta de agregare

plot(ierarhia1)

dm<-as.matrix(d)

View(dm)

#WARD

ierarhia2<-hclust(d,method="ward.D2")

ierarhia2$height

plot(ierarhia2)

#SOLUTIA PENTRU CENTROID

solutia1<-cutree(ierarhia1,k=2)

table(solutia1)

View(solutia1)

rect.hclust(ierarhia1,k=2,border=2:3)

library(cluster)

s<-silhouette(solutia1,d)

plot(s)

s

aggregate(m,list(solutia1),mean)

library(factoextra)

fviz\_cluster(list(data=m,cluster=solutia1))

km2<-kmeans(m,2)

solutiekm2<-km2$cluster

skm<-silhouette(solutiekm2,d)

skm

plot(skm)

#SOLUTIA PENTRU WARD

solutia2<-cutree(ierarhia2,k=2)

table(solutia2)

View(solutia2)

plot(ierarhia2)

rect.hclust(ierarhia2,k=2,border=2:3)

s2<-silhouette(solutia2,d)

plot(s2)

s2

aggregate(m,list(solutia2),mean)

fviz\_cluster(list(data=m,cluster=solutia2))

km2ward<-kmeans(m,2)

solutiekm2ward<-km2ward$cluster

skmWard<-silhouette(solutiekm2ward,d)

skmWard

plot(skmWard)#E LA FEL

#SOLUTIA PENTRU WARD CU 3 CLASE

solutia3<-cutree(ierarhia2,k=3)

table(solutia3)

View(solutia3)

plot(ierarhia2)

rect.hclust(ierarhia2,k=3,border=2:4)

s3<-silhouette(solutia3,d)

plot(s3) #0.49

s3

aggregate(m,list(solutia3),mean)

fviz\_cluster(list(data=m,cluster=solutia3))

km3ward<-kmeans(m,3)

solutiekm3ward<-km3ward$cluster

skmWard3<-silhouette(solutiekm3ward,d)

skmWard3

plot(skmWard3) #0.48